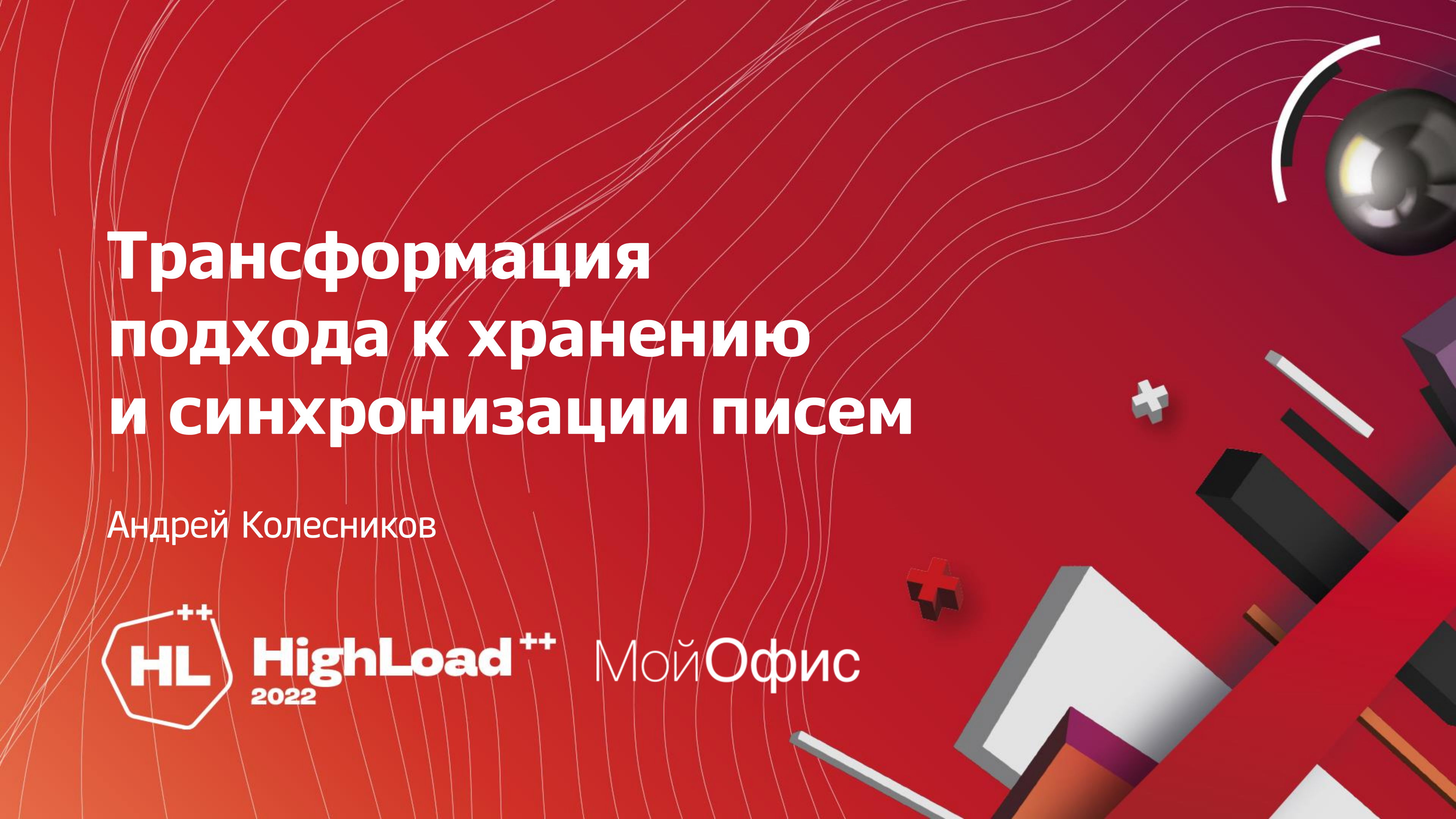

Трансформация
подхода к хранению
и синхронизации писем
Андрей Колесников
HL
HighLoad++
2022
МойОфис

Колесников Андрей

Руководитель инженерного отдела
МойОфис







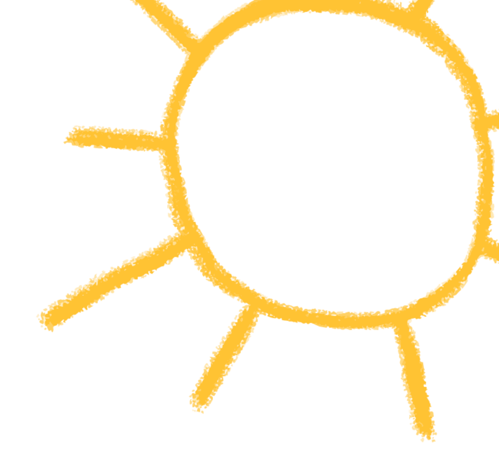













МойОфис

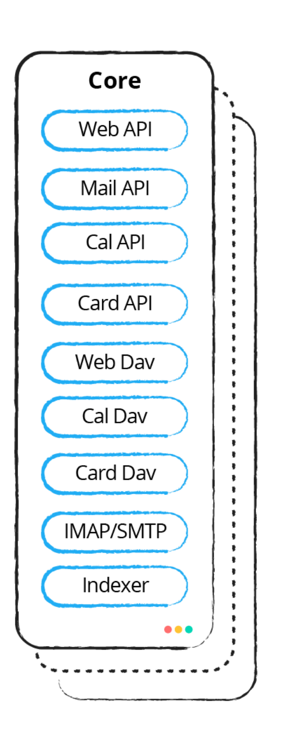

МойОфис

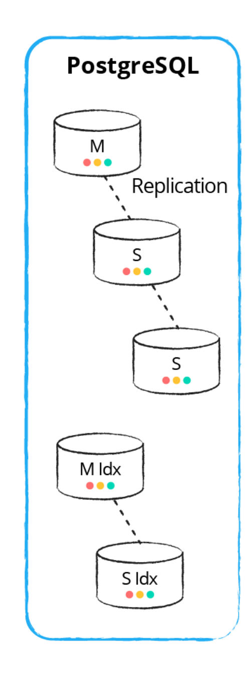

МойОфис

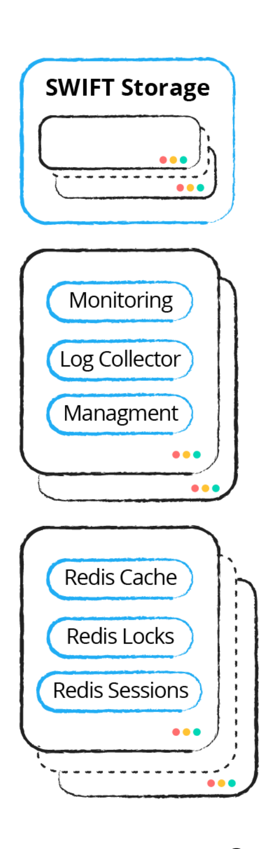

МойОфис

Начало: PostgreSQL, OpenStack Swift

- Метаданные хранятся в PostgreSQL (конверт), тело — в OpenStack Swift
- Кластеризация Pacemaker (менеджер ресурсов) + Corosync (транспортный уровень)









Начало: PostgreSQL, OpenStack Swift

Посмотреть заголовок в базе:

```
SELECT * FROM objects WHERE props->>'envelope' LIKE '%F36A916005E@myoffice.ru%';
----------------------------------------------------------------------------------------------
id          | Oalfa1hDynrh3tNhluiB
owner       |  Ualfa1TT8sOwtqAoMo6
...
title       | Совещание:1 – тема сообщения
...
p_file      | Falfa1hDynrh6axzBS8J
```



Начало: PostgreSQL, OpenStack Swift

Посмотреть заголовок в базе:

```
SELECT * FROM objects WHERE props->>'envelope' LIKE '%F36A916005E@myoffice.ru%';
-----------------------------------------------------------------------------------------
id          | Oalfa1hDynrh3tNhluiB
owner       |  Ualfa1TT8sOwtqAoMo6
...
title       | Совещание:1 – тема сообщения
...
p_file      | Falfa1hDynrh6axzBS8J
```

Получить файл письма:

```
swift -A http://10.2.3.4:8280/auth/v1.0 -U system:syncacc -K XXXXXXXXXXX download
Ualfa1TT8sOwtqAoMo6 Falfa1hDynrh6axzBS8J
```



Начало: PostgreSQL, OpenStack Swift

Посмотреть заголовок в базе:

```
SELECT * FROM objects WHERE props->>'envelope' LIKE '%F36A916005E@myoffice.ru%';
----------------------------------------------------------------------------------------
id          | Oalfa1hDynrh3tNhluiB
owner       |  Ualfa1TT8sOwtqAoMo6
...
title       | Совещание:1 – тема сообщения
...
p_file      | Falfa1hDynrh6axzBS8J
```

Получить файл письма:

```
swift -A http://10.2.3.4:8280/auth/v1.0 -U system:syncacc -K XXXXXXXXXXX download
Ualfa1TT8sOwtqAoMo6 Falfa1hDynrh6axzBS8J
```

```
more Falfa1hDynrh6axzBS8J
```

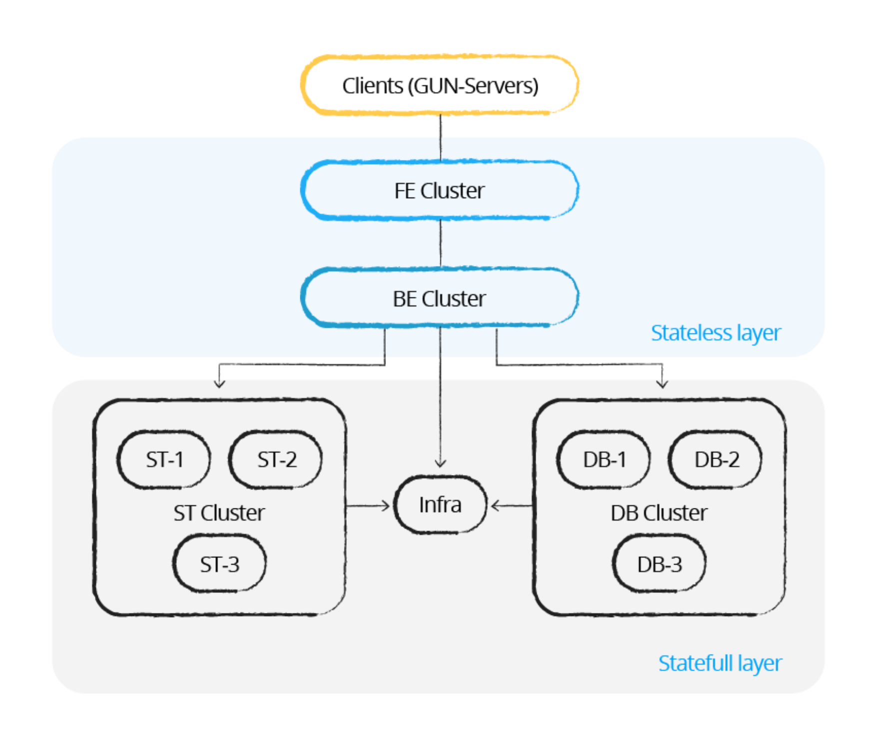


Предел производительности

Максимальная производительность
2500 RPS (100 000 пользователей)

За время тестирования в системе было создано:

- 342 073 пользователя
- 13 833 934 директории
- 18 628 870 файлов



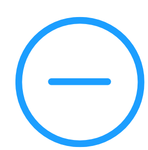


PostgreSQL, OpenStack Swift



Плюсы:

- Наличие экспертизы
- Проверенные временем технологии
- Большое комьюнити



Минусы:

- Архитектурные погрешности, монолит
- Ограничения масштабирования (PostgreSQL по умолчанию поставляется для работы с одним сервером, обслуживающим запросы и двумя резервными)
- Подсистема управления блокировками и кэшированием (Redis) — узкое место для большинства методов



Dovecot dsync

Node 1

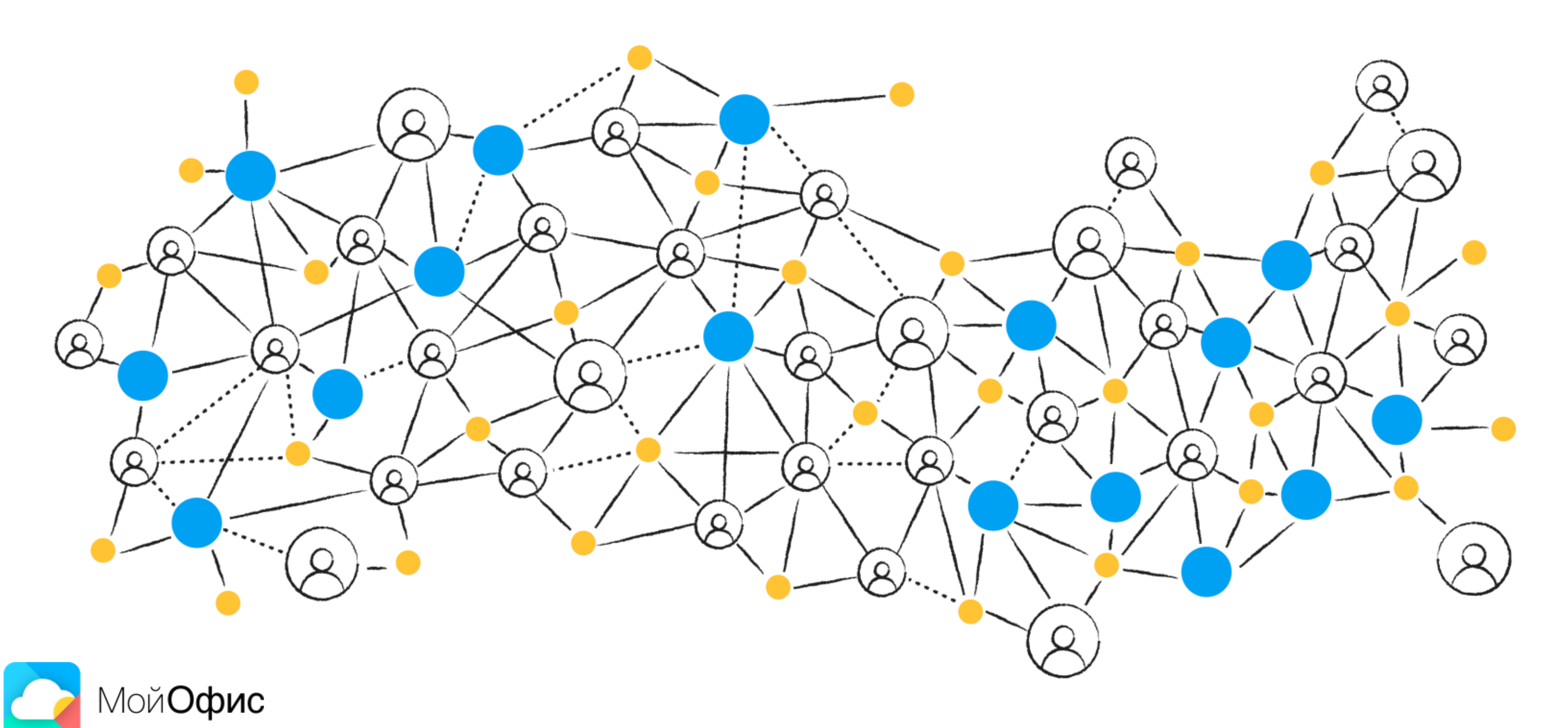
Node 2

Client



Dovecot dsync & Shared Mailboxes

- Расшаренные ящики == проблемы синхронизации dsync: https://wiki.dovecot.org/SharedMailboxes/ClusterSetup



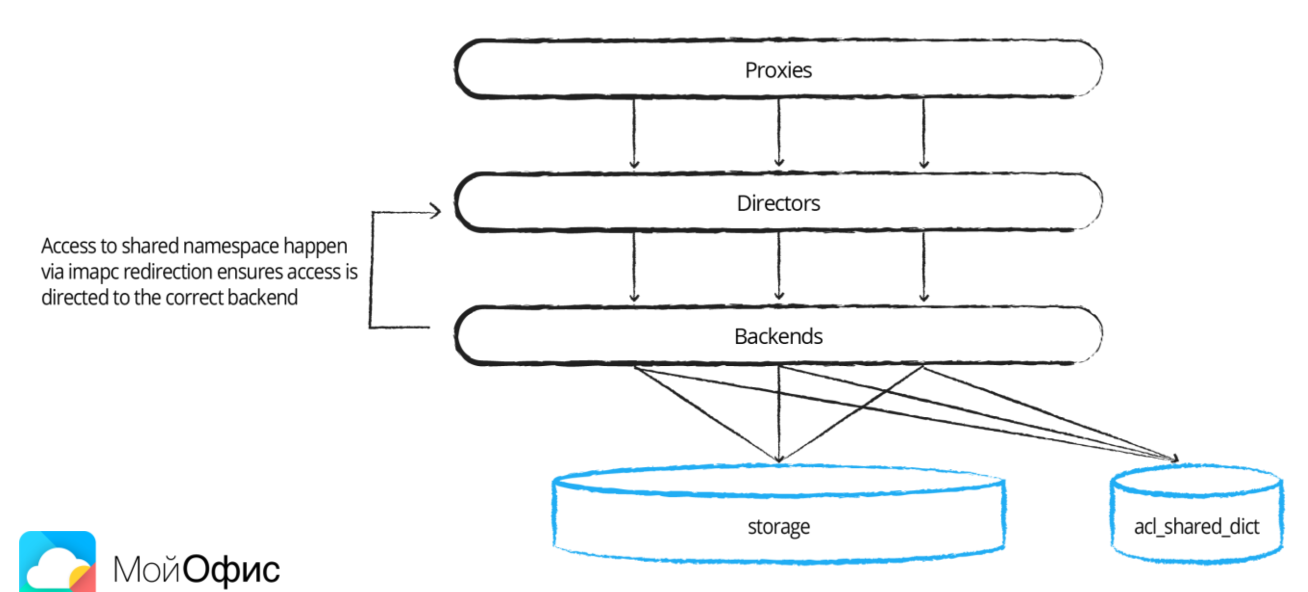
МойОфис

Dovecot dsync: incident resolved

https://wiki.dovecot.org/SharedMailboxes/ClusterSetup

There are some limitations for this kind of use case:

- `imapc_*` settings are global. You can't have two different namespaces with different imapc setting yet.
- The imapc code doesn't support some IMAP features. Most importantly SORT isn't supported, which may result in lower performance.



МойОфис

GlusterFS

Brick – каталог хранения, основная единица хранения GlusterFS

Volume – том, логическая комбинация каталогов хранилища

```
gluster volume info
```

```
Volume Name: gv0
Type: Replicate
Volume ID: f25cc3d8-631f-41bd-96e1-3e22a4c6f71f
Status: Started
Snapshot Count: 0
Number of Bricks: 1 x 3 = 3
Transport-type: tcp
Bricks:
Brick1: server1:/data/brick1/gv0
Brick2: server2:/data/brick1/gv0
Brick3: server3:/data/brick1/gv0
Options Reconfigured:
transport.address-family: inet
```



GlusterFS

Replicated — тома с репликацией. Аналогично RAID 1. В такой конфигурации одни и те же данные записываются минимум на два подтома.

Mount Point

Replicated Volume

Server1:exp1/

Brick

Server2:exp2/

Brick

File1

File2



GlusterFS

./hello /tmp/fuse

libfuse

glibc

ls - l /tmp/fuse

glibc

Userspace

Kernel

VFS

FUSE

NFS

Ext3

...

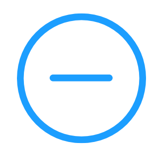


GlusterFS



Плюсы:

- Быстрый старт
- Большое комьюнити
  и сопровождение Red Hat
  (В 2011 проект был приобретен
  компанией Red Hat и лег в основу
  Red Hat Storage Server

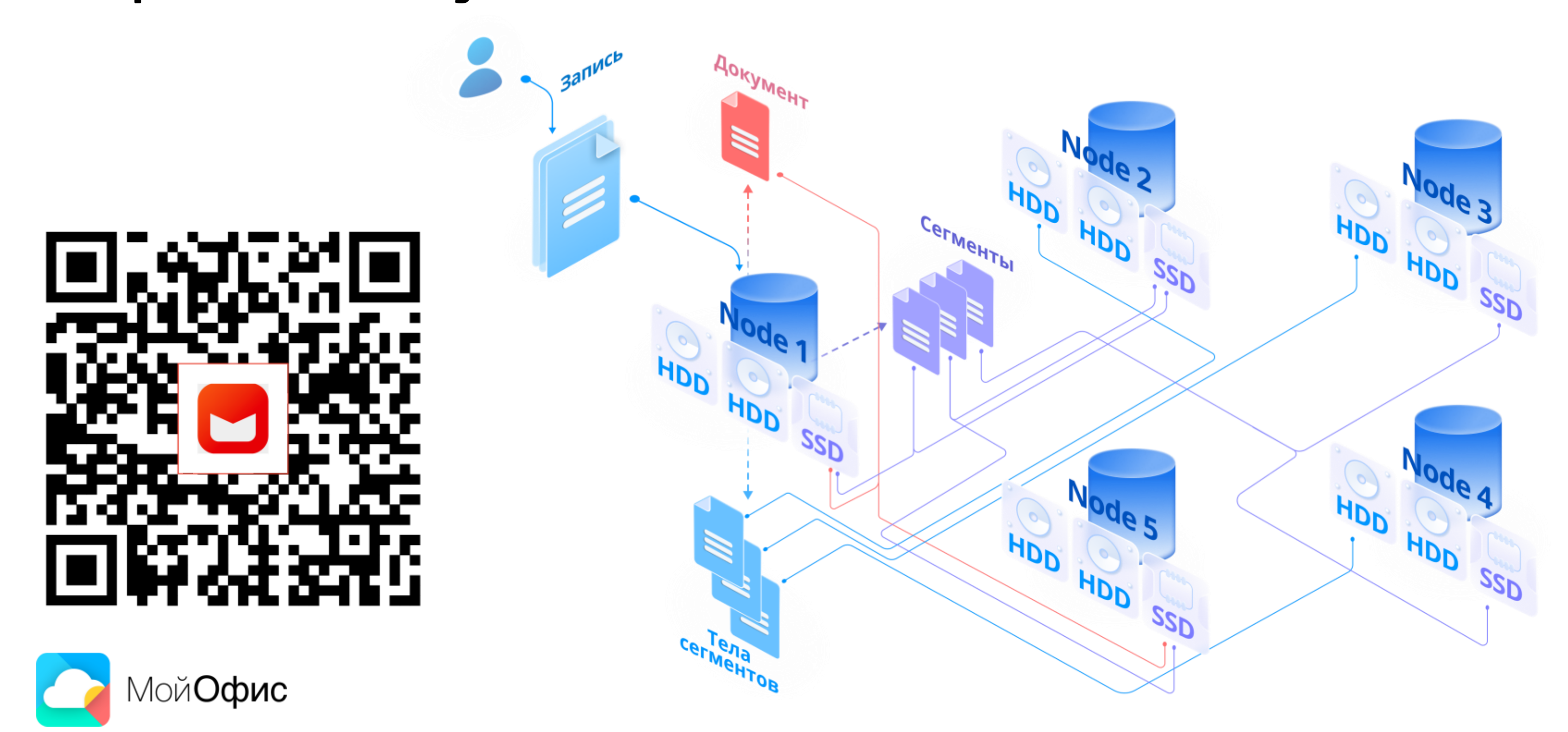


Минусы:

- Чувствительность к задержкам
  в сети (> 50 ms)
- Массовое удаление файлов:
  Remote I/O error

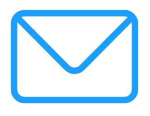


Собственная разработка:
Dispersed Object Store



МойОфис

Особенности корпоративной переписки



Большое количество полностью или частично совпадающих писем:

- У каждого письма минимум 2 копии (отправитель и получатель).
- Корпоративные рассылки (ещё больше копий).
- Шаблонные письма (ERP-системы, баг-трекеры, календари).
- Повторяющиеся элементы писем (вложения, подписи).
- Длинные цепочки писем со взаимным цитированием.



Структура электронного письма

Заголовок письма ▶

От: sergey.sergeev@myoffice.team
Кому: alexey.pogudin@myoffice.team
Дата: 23.09.2022

{
"from":"sergey.sergeev@myoffice.team",
"to": "alexey.pogudin@myoffice.team",
"created_at":"2020-09-23T10:01:02Z+03:00"
}

Части письма

Основной текст ▶

Здравствуйте, отправляю отчёт
о взаимодействии с корпоративными
клиентами за I квартал 2022 года.

Сергей Сергеев

Руководитель направления
по работе с корпоративными клиентами

Здравствуйте, отправляю отчёт
о взаимодействии с корпоративными
клиентами за III квартал 2020 года.

Сергей Сергеев

Руководитель направления
по работе с корпоративными клиентами

Картинка из подписи ▶

МойОфис

МойОфис
logo.png

Вложенный файл ▶

Вложения: PDF report.pdf

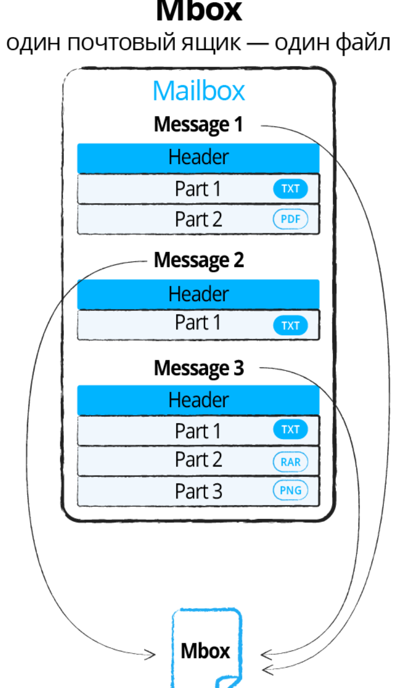
PDF
report.pdf



Способы хранения электронной почты

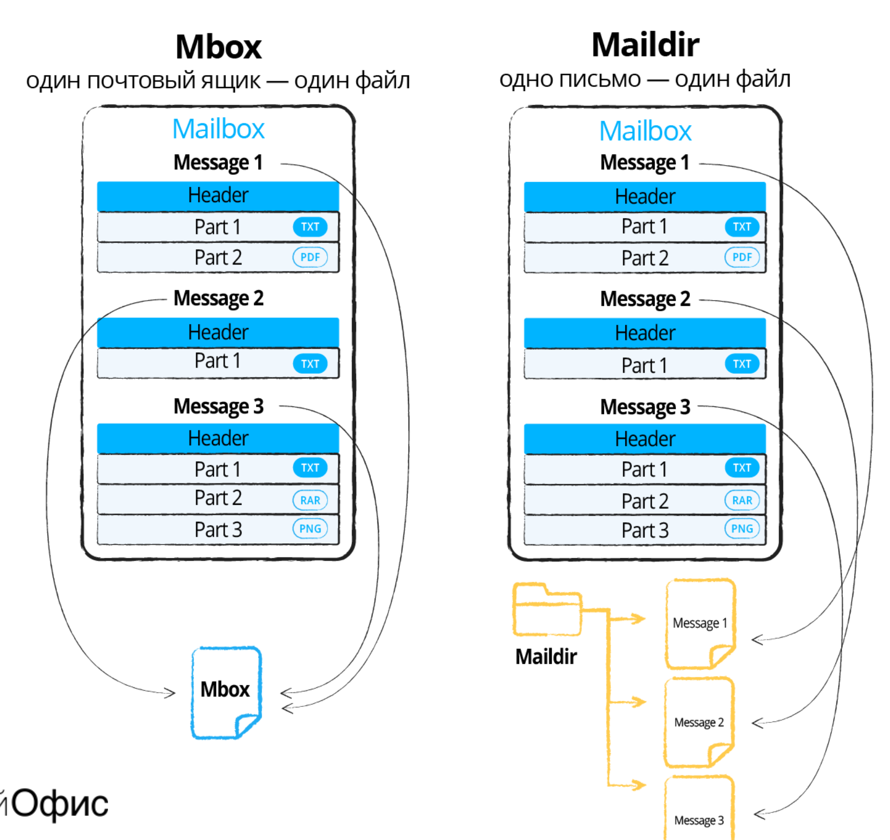


один почтовый ящик — один файл



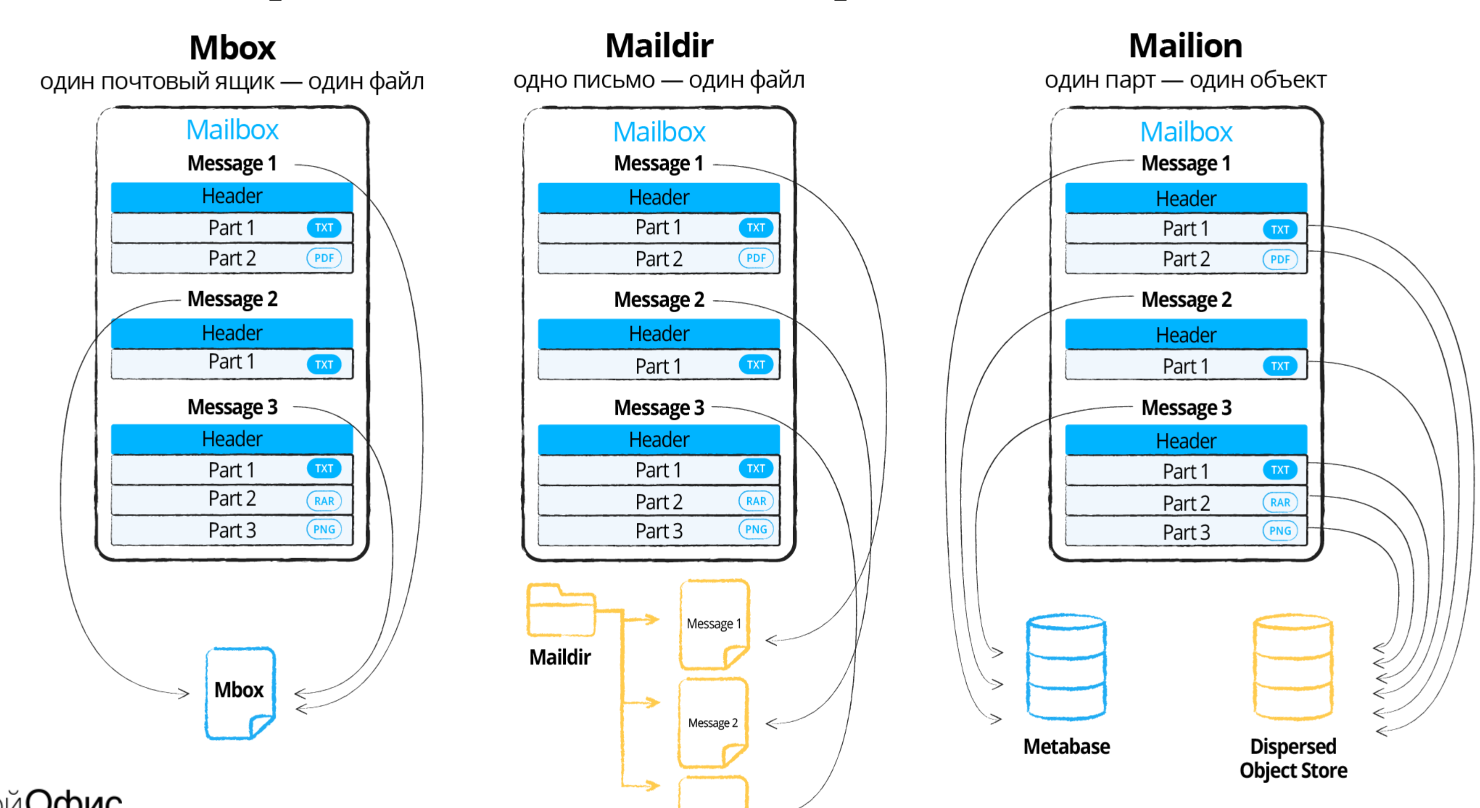


Способы хранения электронной почты



МойОфис

Способы хранения электронной почты



МойОфис

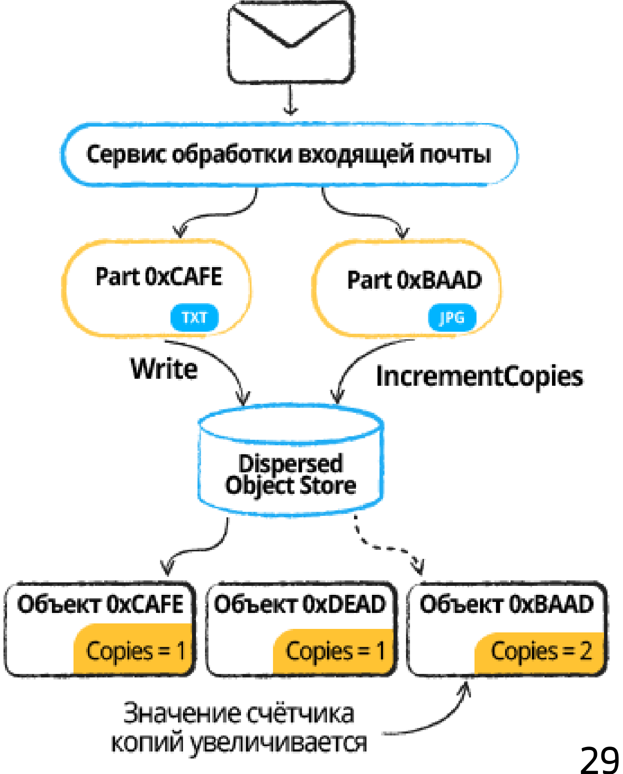
Дедупликация на уровне объектов





Дедупликация на уровне объектов







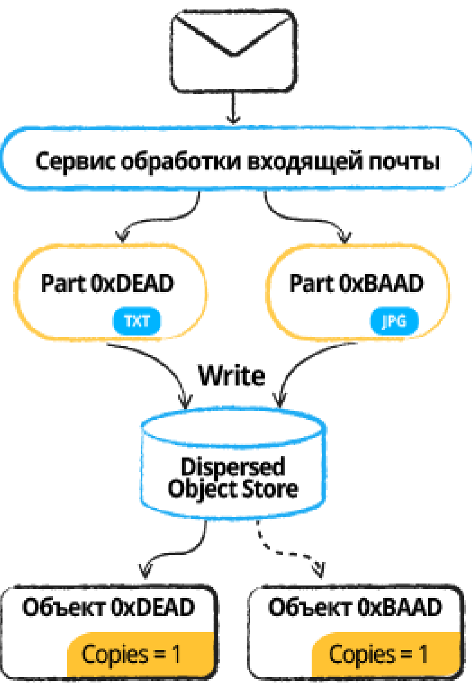
Дедупликация на уровне объектов

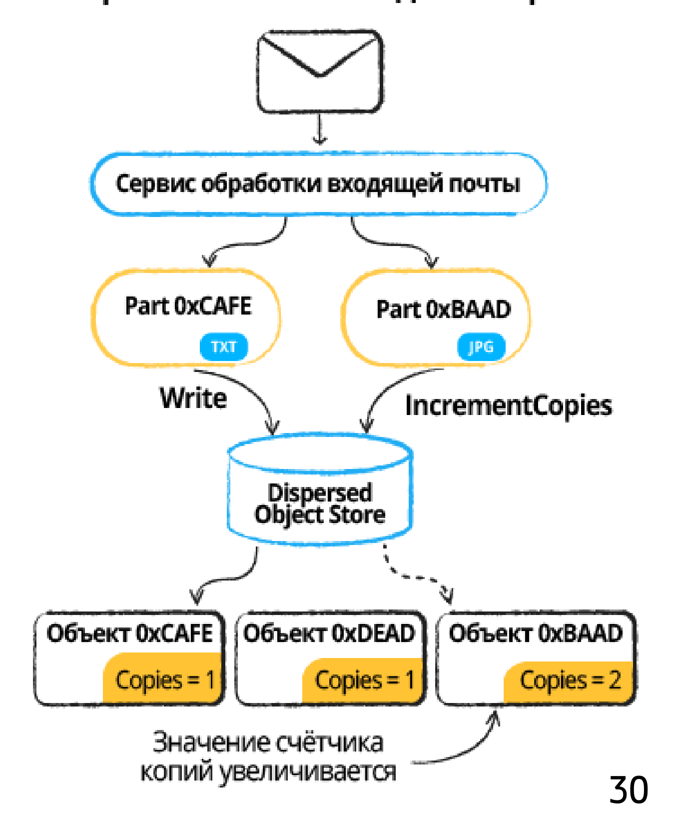


Плюсы:
- Экономия дискового пространства
- Экономия сетевого трафика



Минусы:
- Усложнение логики клиентского сервиса







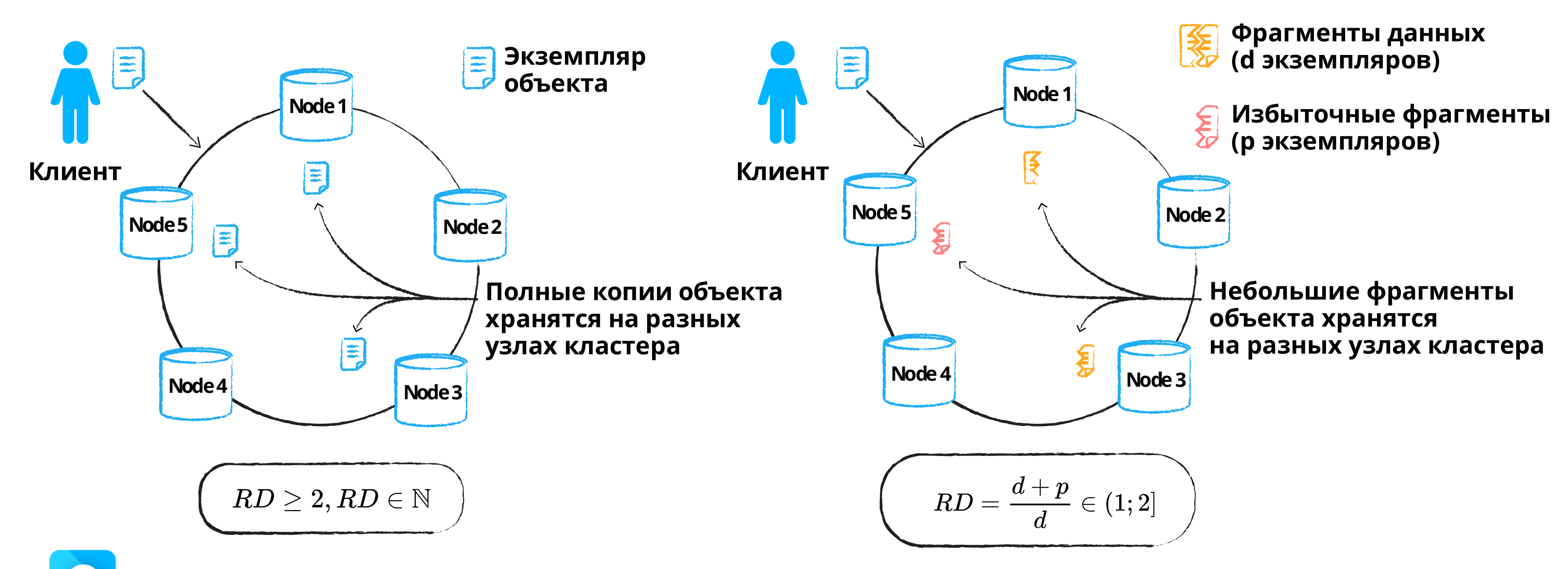
Тип и размер данных

- Хранилище умеет определять формат поступивших данных:
  - ✓ Text (txt, html, rfc822)
  - ✓ Binary (image, video, audio, pdf)
- Разные конвейеры обработки для разных типов данных.
- Размер данных имеет значение.

Класс размеров	Текстовые данные		Бинарные данные	
	Чанкинг	Компрессия	Чанкинг	Компрессия
Большие, средние	Content-defined chunking	+	Equal-size chunking	-
Малые	-	-	-	-



Отказоустойчивость и избыточность

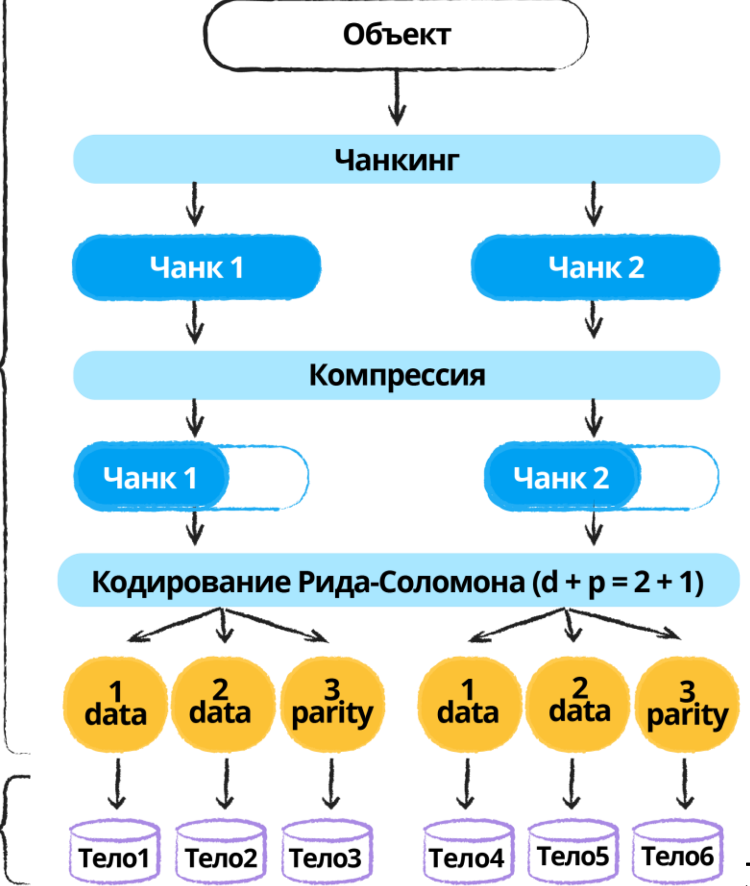


МойОфис

Отказоустойчивость и избыточность

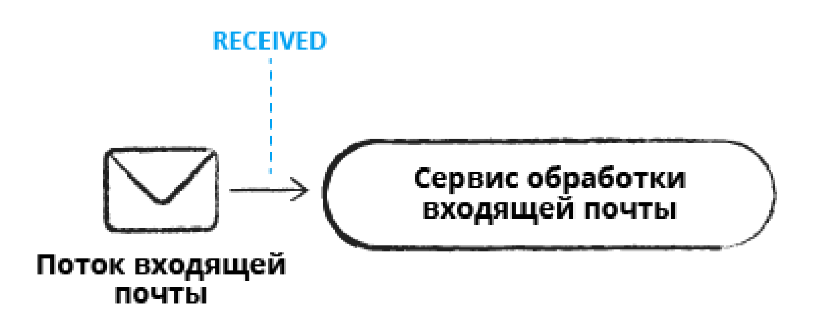
- Соотношение метаданных к данным изменяется в пределах 1:3 — 1:500.
- Для каждого слоя хранения установлен свой уровень избыточности.

$$RD_{meta} \in \{2, 3, 5\}$$
$$RD_{data} \in (1; 2]$$

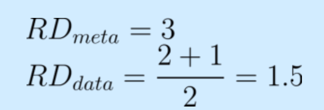




Эффективность оптимизаций



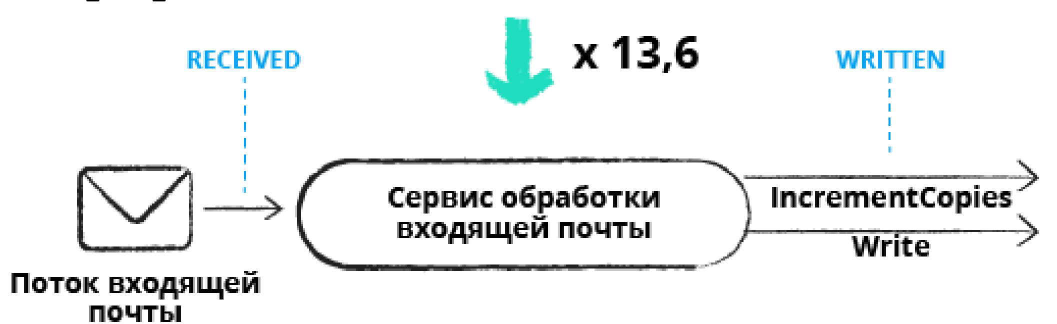




$$RD_{meta} = 3$$
$$RD_{data} = \frac{2+1}{2} = 1.5$$



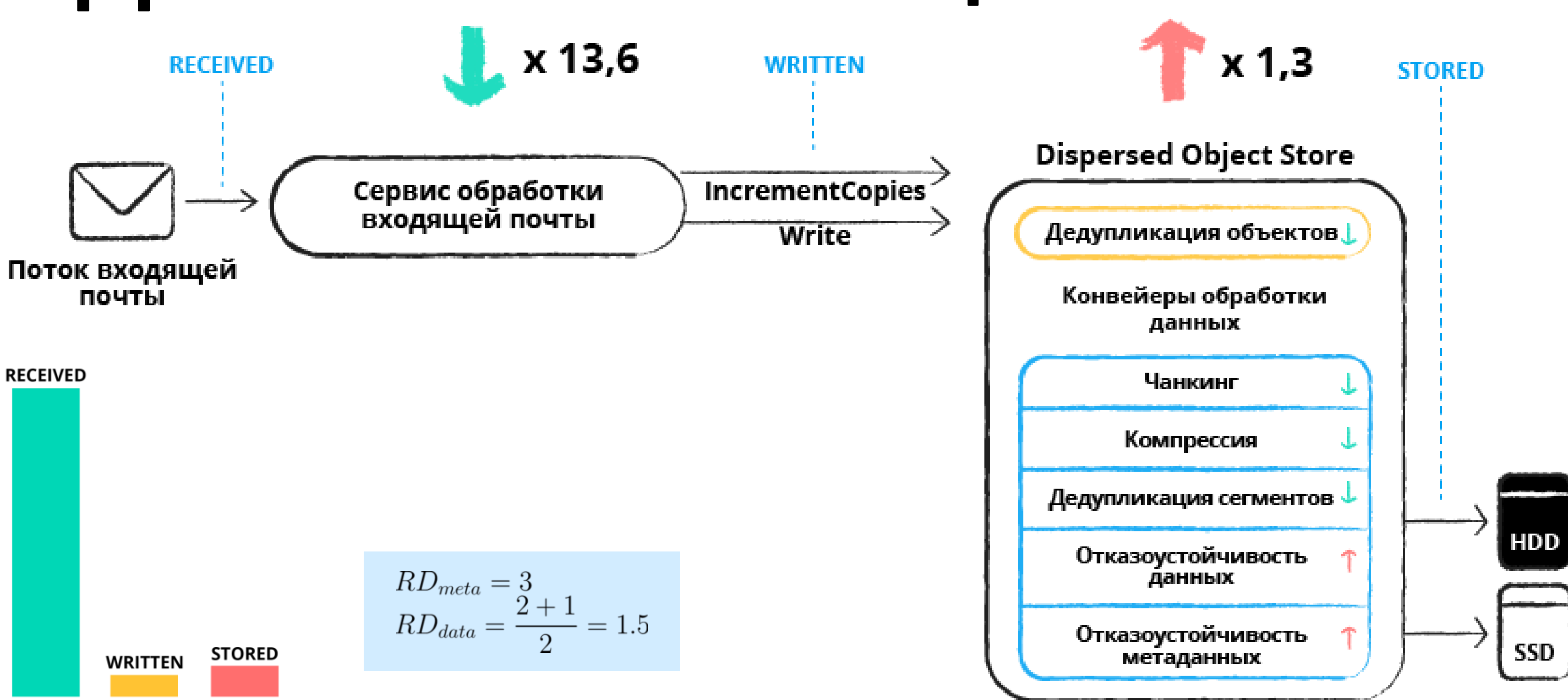
Эффективность оптимизаций





$$RD_{meta} = 3$$
$$RD_{data} = \frac{2+1}{2} = 1.5$$

Эффективность оптимизаций



$$RD_{meta} = 3$$
$$RD_{data} = \frac{2+1}{2} = 1.5$$

Эффективность оптимизаций

x 13,6

RECEIVED

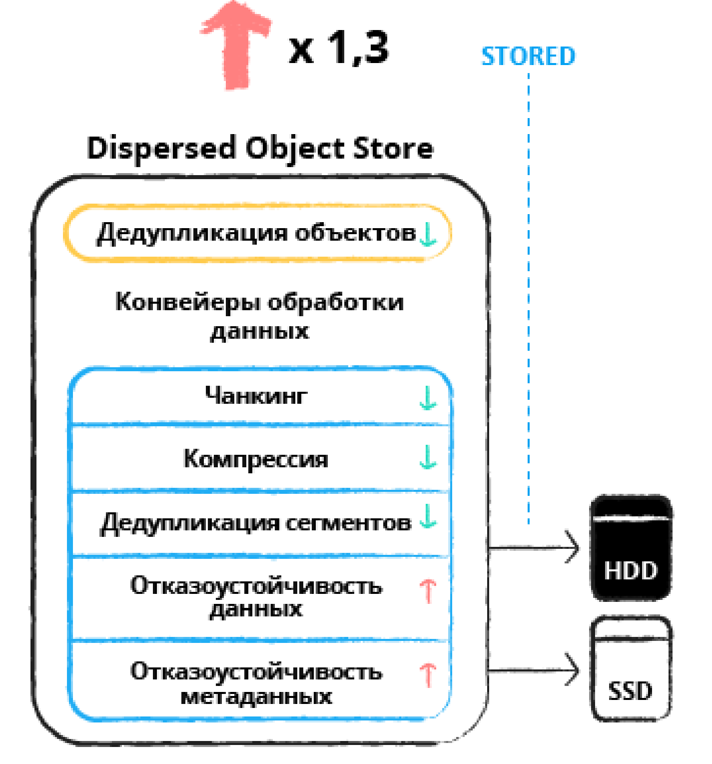
WRITTEN

Поток входящей почты

Сервис обработки входящей почты

IncrementCopies

Write



RECEIVED

WRITTEN

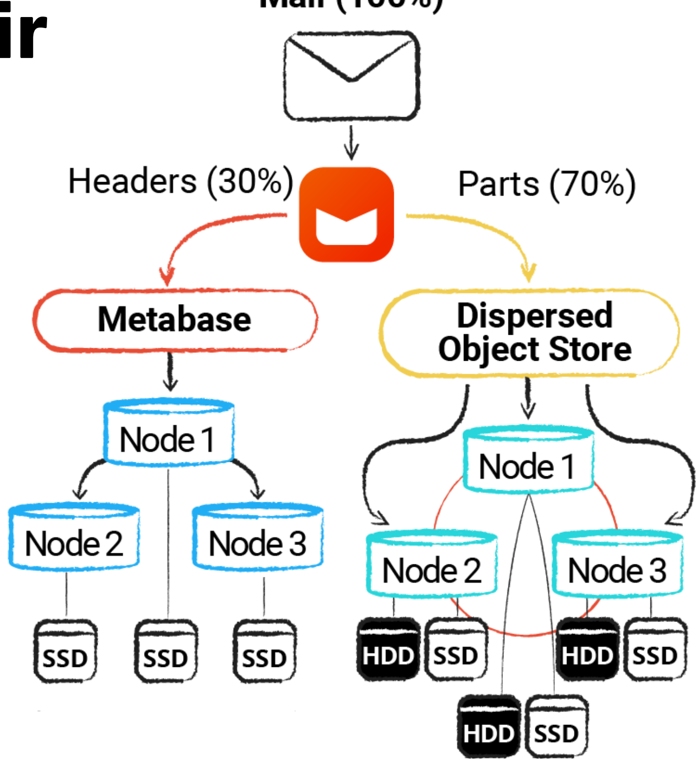
STORED

$$\frac{Received}{Stored} = \frac{Received}{Written} \cdot \frac{Written}{Stored} = \frac{13.6}{1.3} \approx 10.5$$

$$RD_{meta} = 3$$
$$RD_{data} = \frac{2+1}{2} = 1.5$$



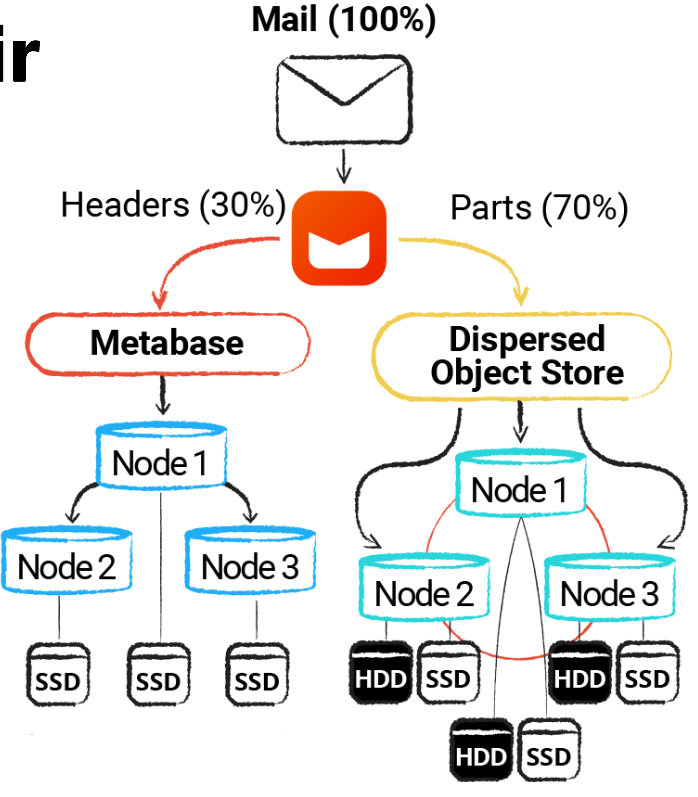
Mailion vs Maildir



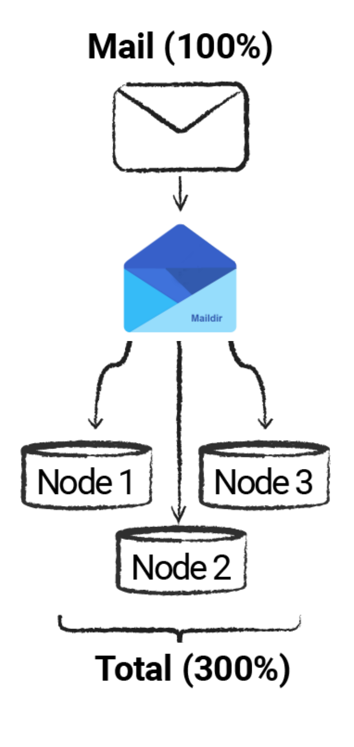
$$RD_{meta} = 3$$
$$RD_{data} = \frac{2+2}{2} = 2$$



Mailion vs Maildir



Total (16.6%)



$$RD_{meta} = 3$$
$$RD_{data} = \frac{2+2}{2} = 2$$

$$RD = 3$$

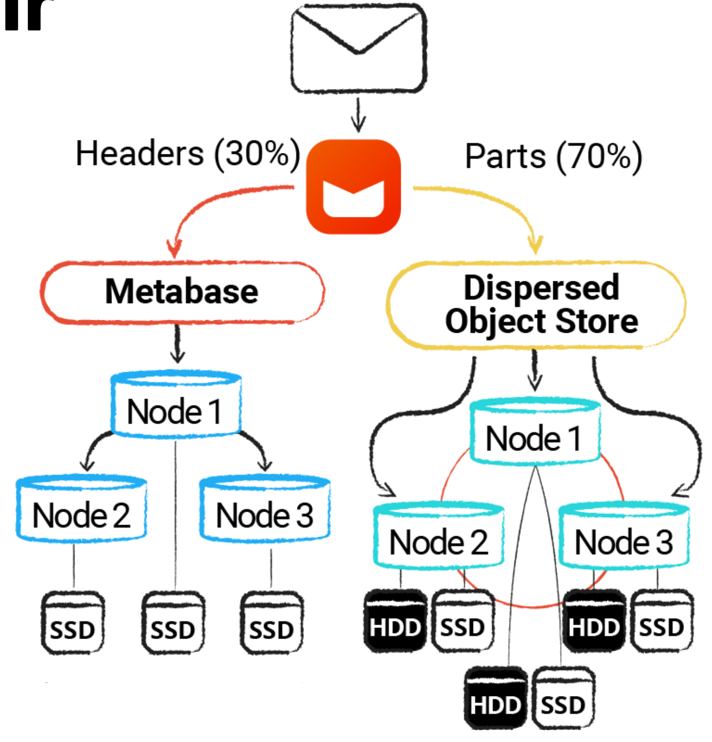


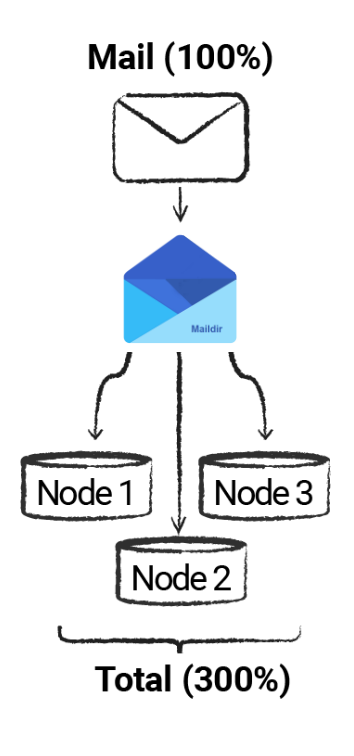
Mailion vs Maildir

- Потребление дискового пространства ниже в 18 раз (300% / 16.6% = 18.07).
- Стоимость хранения данных ниже в 8 раз.

$$RD_{meta} = 3$$
$$RD_{data} = \frac{2+2}{2} = 2$$

$$RD = 3$$







Расчёт стоимости хранения данных

Wikibon HDD & SSD Price-reduction Projections 2020 - 2030												
	2020	2021	2022	2023	2024	2025	2026	2027	2028	2029	2030	CAGR 2020-30
HDD $/TB	$ 22	$ 21	$ 19	$ 18	$ 17	$ 16	$ 15	$ 14	$ 14	$ 13	$ 13	-5.6%
SSD $/TB	$ 128	$ 86	$ 58	$ 40	$ 29	$ 21	$ 15	$ 12	$ 9	$ 7	$ 6	-26.8%
Ratio SSD/HDD	5.8	4.2	3.0	2.2	1.7	1.3	1.0	0.8	0.7	0.54	0.45	-22.5%
HDD Yr % decrease	-7.2%	-6.9%	-6.5%	-6.2%	-5.9%	-5.6%	-5.4%	-5.1%	-4.9%	-4.6%	-4.4%	
SSD Yr % decrease	-34%	-33%	-32%	-31%	-29%	-28%	-26%	-25%	-23%	-22%	-20%	
Ratio SSD/HDD	-29%	28%	27%	26%	25%	23%	22%	20%	19%	18%	16%	
Source: © Wikibon, 2021												

https://wikibon.com/qlc-flash-hamrs-hdd

- **500TB x 3 = 1,5PB (cluster Maildir)**
- **1,5PB x 19$ = 28 500$ (cluster Maildir)**



Расчёт стоимости хранения данных

Wikibon HDD & SSD Price-reduction Projections 2020 - 2030												
	2020	2021	2022	2023	2024	2025	2026	2027	2028	2029	2030	CAGR 2020-30
HDD $/TB	$ 22	$ 21	$ 19	$ 18	$ 17	$ 16	$ 15	$ 14	$ 14	$ 13	$ 13	-5.6%
SSD $/TB	$ 128	$ 86	$ 58	$ 40	$ 29	$ 21	$ 15	$ 12	$ 9	$ 7	$ 6	-26.8%
Ratio SSD/HDD	5.8	4.2	3.0	2.2	1.7	1.3	1.0	0.8	0.7	0.54	0.45	-22.5%
HDD Yr % decrease	-7.2%	-6.9%	-6.5%	-6.2%	-5.9%	-5.6%	-5.4%	-5.1%	-4.9%	-4.6%	-4.4%	
SSD Yr % decrease	-34%	-33%	-32%	-31%	-29%	-28%	-26%	-25%	-23%	-22%	-20%	
Ratio SSD/HDD	-29%	28%	27%	26%	25%	23%	22%	20%	19%	18%	16%	

Source: © Wikibon, 2021

https://wikibon.com/qlc-flash-hamrs-hdd

- **500TB x 3 = 1,5PB (cluster Maildir)**
- **1,5PB x 19$ = 28 500$ (cluster Maildir)**
- **<u>28 500$ / 8 = 3 562,5$ (cluster DOS)</u>**



Нагрузочное тестирование



- Профиль нагрузки на 600 000 пользователей:

  - вход в систему;
  - отправка новых писем;
  - создание событий в календаре и реагирование на них;
  - другое.

- Для тестирования использовали K6 компании Grafana Labs

- Скрипты тестирования запускались на группировке из 46 серверов, которые суммарно были оснащены 636 процессорными ядрами, 2,8 ТБ оперативной памяти и накопителями емкостью более 135 ТБ.



Нагрузочное тестирование

- В ходе испытаний инженерами проверялась гипотеза о стабильной работе системы под нагрузкой **6081 RPS** (операций в секунду). Это эквивалентно действиям **600 000 пользователей**, которые в течение дня отправляют и получают **1,14 млн писем**.

Requests per Second

15 K
10 K
5 K
0

16:30 17:00 17:30 18:00 18:30 19:00 19:30 20:00 20:30 21:00 21:30 22:00 22:30 23:00 23:30 00:00

Requests per Second Avg: 6.53 K



Нагрузочное тестирование: 6081 RPS

Распределение нагрузки на DOS

140
120
100
80
60
40
20
0

время ответа p99, мс

58% запросы на изменение счётчика копий у данных
28% запросы на запись данных
14% запросы на чтение данных



Нагрузочное тестирование: раунд 1

- Выявлены критические задержки некоторых методов

Errors Per Second (by Methods)

550
500
450
400
350
300
250
200
150
100
50
0

16:30 17:00 17:30 18:00 18:30 19:00 19:30 20:00 20:30



Нагрузочное тестирование: раунд 2

- Корректировка конфигурации (подключения к базе, кэш)

Errors Per Second (by Methods)

1
0.800
0.600
0.400
0.200
0
-0.20
-0.40
-0.60
-0.80
-1

04:10 04:20 04:30 04:40 04:50 05:00 05:10 05:20 05:30 05:40 05:50 06:00 06:10 06:20 06:30 06:40 06:50 07:00 07:10 07:20 07:30 07:40 07:50 08:00 08:10 08:20 08:30 08:40 08:50 09:00 09:10 09:20 09:30 09:40 09:50 10:00 10:10 10:20 10:30 10:40 10:50 11:00 11:10 11:20 11:30

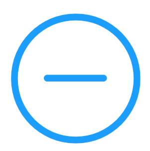


Собственное объектное хранилище



Плюсы:

- Контроль над разработкой, приоритизацией
- Дедупликация и компрессия
- Повышение экономической эффективности почтовой системы
- Оптимальный баланс аппаратных ресурсов (CPU / RAM vs Disk / IO)



Минусы:

- Сложная и дорогая разработка (распределённый stateful-сервис)




МойОфис



Андрей Колесников

8-929-384-19-24

av_kolesnikov



Оценить доклад



Слайды: https://bit.ly/3sM0ihW